

AVALIANDO TÉCNICAS DE NORMALIZAÇÃO PARA MICROARRAYS DE cDNA

Fernando Henrique Ferraz Pereira da Rosa
Instituto de Matemática e Estatística
Universidade de São Paulo
feferraz@ime.usp.br

Júlia Maria Pavan Soler
Instituto de Matemática e Estatística
Universidade de São Paulo
pavan@ime.usp.br

Resumo

Estudos em genética envolvendo experimentos com microarrays permitem a quantificação e a comparação simultânea dos níveis de expressão de genes em uma larga escala. A técnica empregada nesses estudos entretanto possui uma variabilidade considerável [1], fazendo-se necessário um ajuste preliminar dos dados antes que possa ser analisada a variabilidade biológica, na qual geralmente nosso interesse está focado. A esse ajuste preliminar dos dados antes da análise inferencial se dá o nome de *normalização* [2].

Nesse trabalho introduzimos os principais modelos de normalizações ([3],[4]) disponíveis na literatura, estudando suas propriedades e comparando seus resultados.

A maioria desses métodos de normalização se baseia em métodos de alisamento do gráfico MA das intensidades de uma dada lâmina de microarray, para obtenção das constantes de correção. O método mais consagrado na literatura é o uso de regressão robusta local, *lowess*, que apresenta bons resultados na correção da variabilidade da técnica.

Depois de estudar os métodos disponíveis na literatura, propomos também um novo método de normalização utilizando *smoothing splines*, e comparamos seu desempenho com os métodos até então disponíveis. Os resultados mostram que a maior adaptatividade da regressão por splines consegue melhores normalizações, controlando melhor a variabilidade devida à técnica.

Introdução

Um experimento com microarrays de cDNA típico envolve a hibridização de duas amostras de mRNA, provenientes de um grupo controle e de um grupo experimental, que são convertidas em cDNA e marcadas com corantes fluorescentes. A partir dessas marcações, são então geradas leituras de intensidade para cada canal fluorescente, que fornecem informações sobre a expressão relativa dos genes inclusos na lâmina de microarray.

O objetivo desse tipo de estudo, em geral, consiste na identificação de genes diferentemente expressos entre os dois grupos. Essa identificação entretanto não pode ser realizada comparando-se diretamente os sinais de intensidades obtidos para cada gene. Isso ocorre pelo fato de haver uma grande variabilidade da técnica, devido a fatores como diferente peso atômico dos corantes fluorescentes utilizados (acarretando uma maior taxa de fixação para um dos corantes), variação das taxas de hibridização devido à localização espacial dos spots na lâmina, eventuais falhas na hibridização, estouro das leituras de intensidade, entre outros.

Como uma forma de contornar esses problemas, foram desenvolvidas diversas técnicas de normalização ([2], [4], [5]), que objetivam controlar as variações sistemáticas entre os níveis de intensidade medidos em uma co-hibridização, de forma que a variação *biológica* se sobressaia à variação da *técnica* e os genes diferentemente expressos possam ser identificados.

Dados

O conjunto de dados utilizado nas análises é proveniente de um experimento realizado no Laboratório de Cardiologia Molecular do Instituto do Coração (InCor-USP) com ratos congênicos no qual o fenótipo de interesse é a hipertensão. Foram utilizadas lâminas de vidro com superfície de polyisina, com 26912 spots cada (cada spot pode conter uma seqüência de cDNA que pode corresponder a um gene, ESTs ou seqüências de controle).

Para cada spot de cada hibridização há leituras de intensidade para dois canais, o verde (G/green) e o vermelho (R/red). Há portanto um conjunto de 26912 pares de leituras de intensidade (r, g) .

Nesse estudo, o interesse está focado na identificação da variabilidade biológica devida a fatores genéticos, pois os animais foram obtidos através de cruzamentos de linhagens congênicas em laboratório, de forma que outros fatores que poderiam ter influência foram controlados.

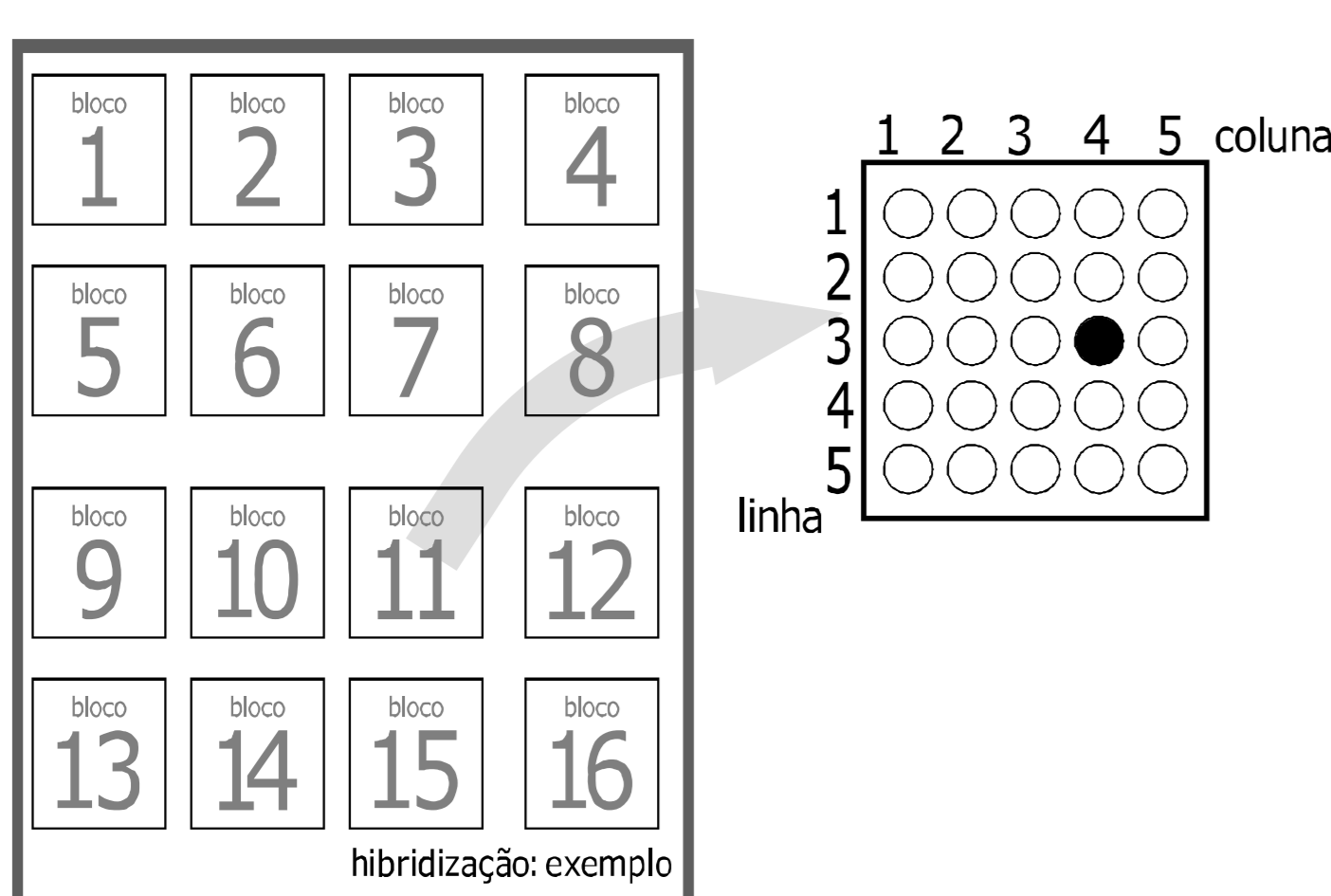


Figura 1: Localização de um spot numa lâmina de microarrays fictícia

Hibridização	Cy3	Cy5
rat85	2c-controle	SHR-controle
rat89	SHR-controle	2c-controle
rat88	2c-sal	SHR-sal
rat86	SHR-sal	2c-sal

Tabela 1: Delineamento Experimental

Gráfico MA

Os dados de intensidade de fluorescência são tradicionalmente analisados utilizando-se uma transformação dos dados iniciais. Toma-se $M = \log_2 R/G$ e $A = \log_2 \sqrt{RG}$, dando origem ao MAPlot, que é o diagrama de dispersão de M por A . Esse gráfico nada mais é do que um tipo de gráfico de média por desvio-padrão, apropriado em geral para identificação espacial de pontos observados.

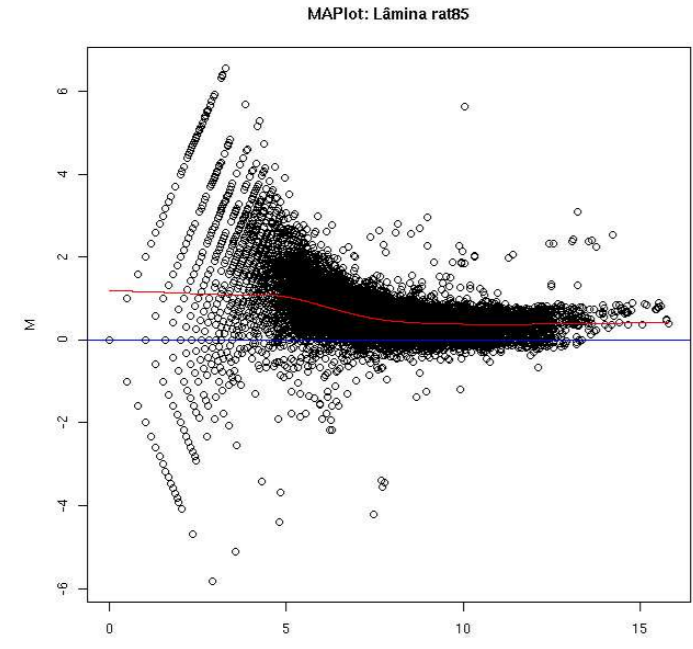


Figura 2: MAPlot para lâmina rat85

Normalização

Sobre os dados transformados para a escala M-A, são aplicadas as técnicas de normalização, que buscam reduzir a variação da técnica e tornar as intensidades comparáveis entre hibridizações diferentes. Dado um array com s spots, toma-se a transformação:

$$M_k \rightarrow M_k - C(A_k, M_k), \quad k = 1, \dots, s$$

onde $C(A_k, M_k)$ é uma dada função normalizadora adequada.

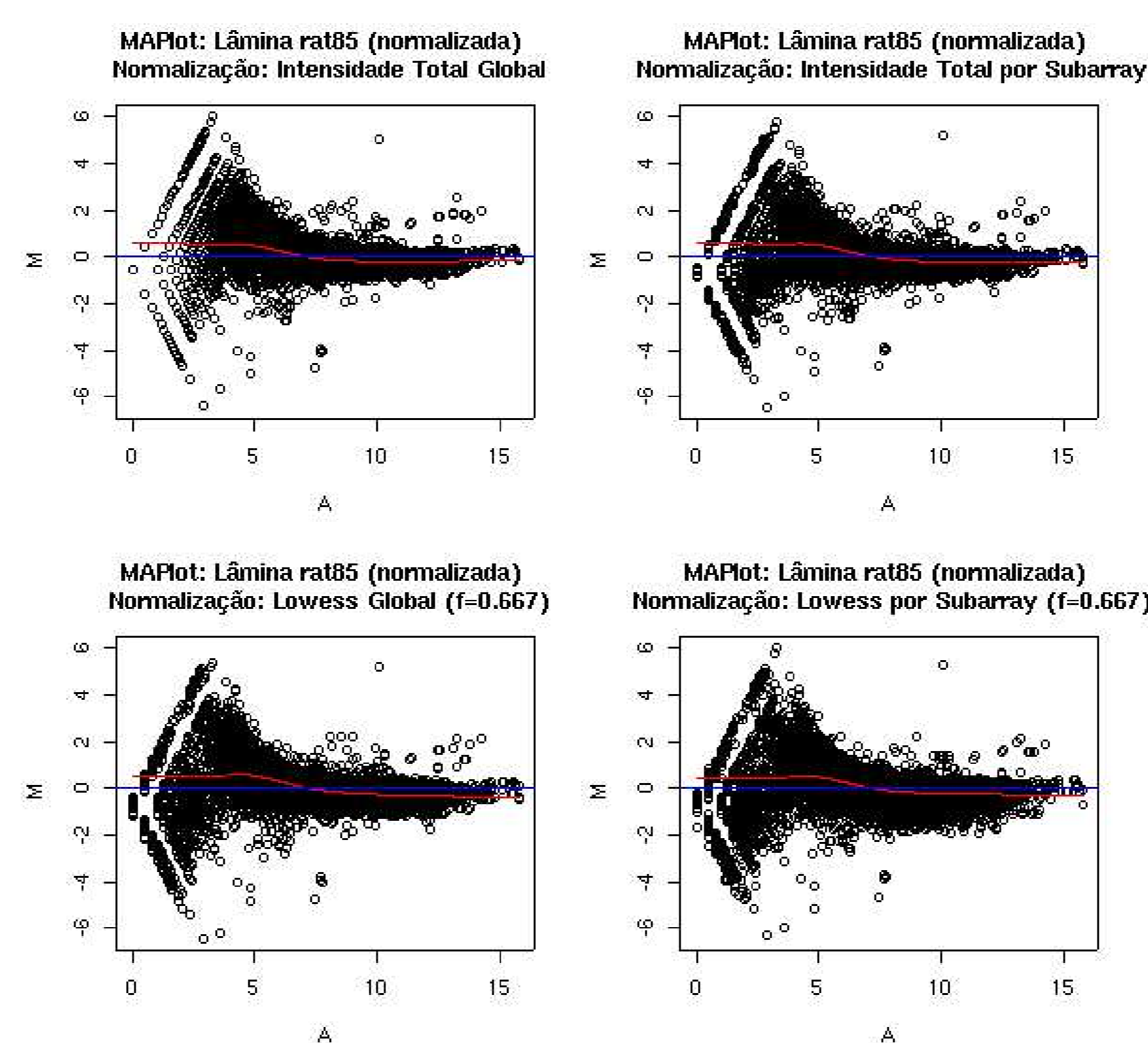


Figura 3: Diferentes normalizações comumente utilizadas

Normalização por smoothing splines

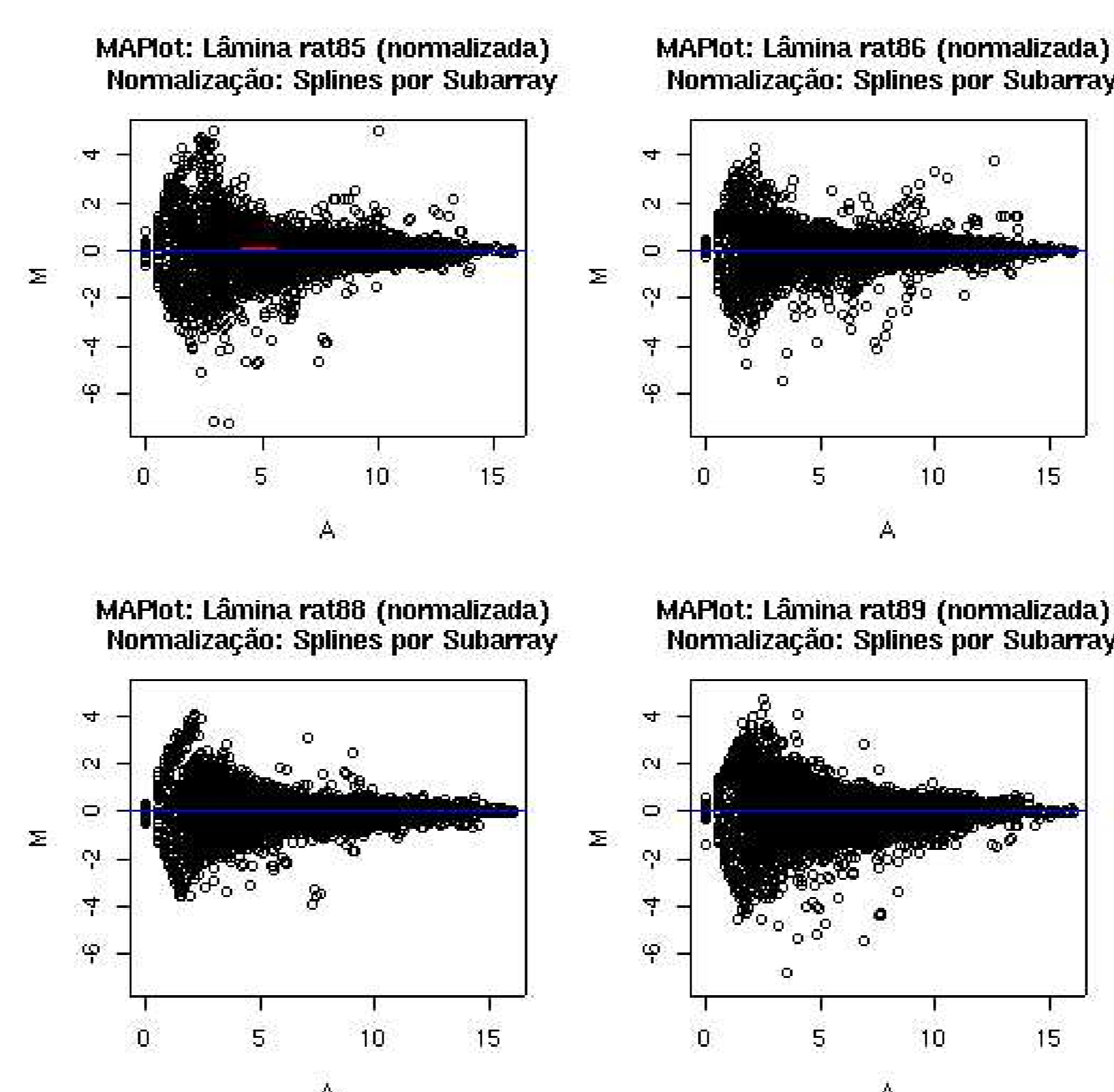


Figura 4: Normalizações por smoothing splines

Auto-normalização

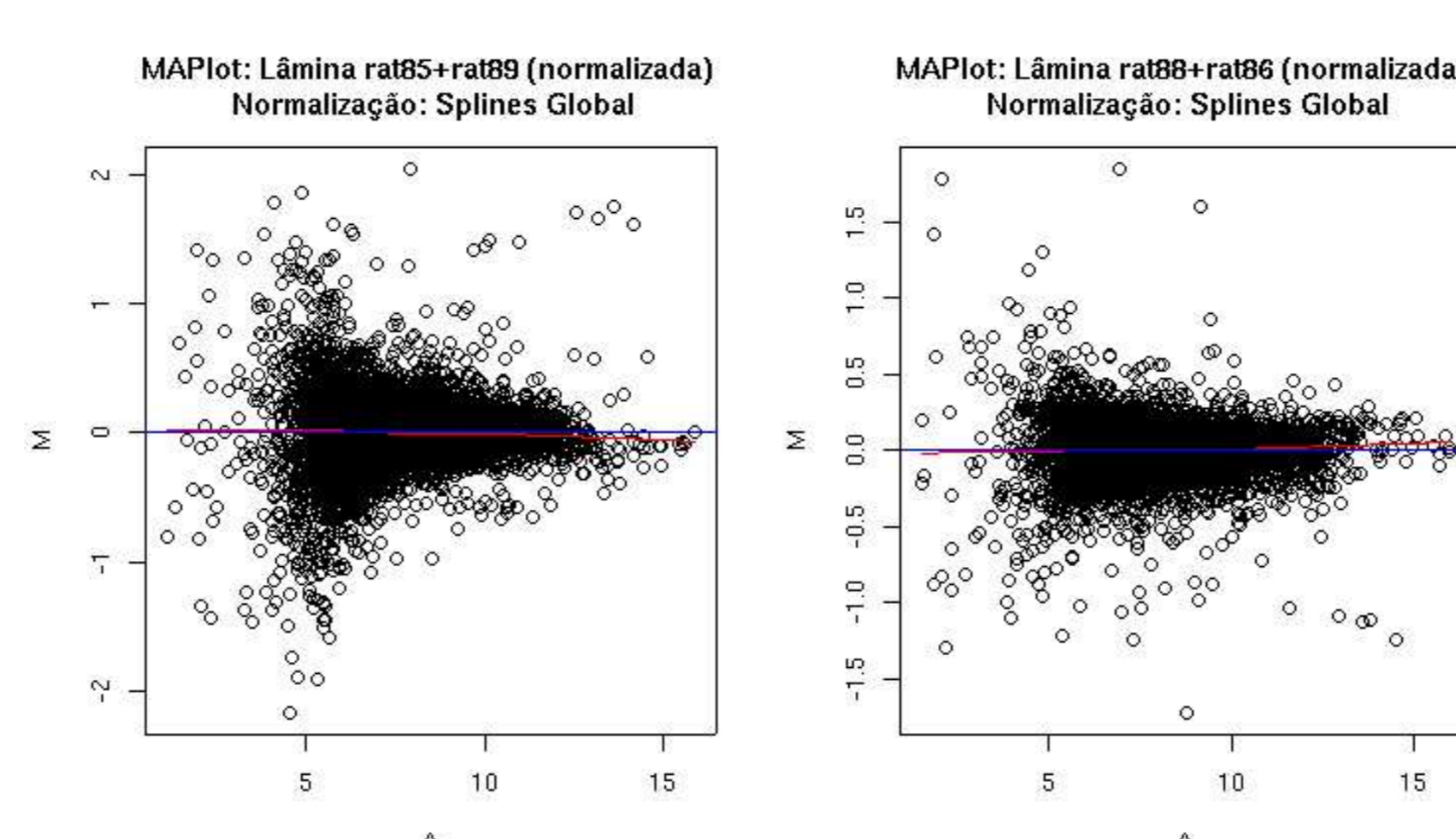


Figura 5: Auto-normalização
rat85+rat89: Grupo controle
rat86+rat88: Grupo sal

Medidas de Comparação

Definimos e aplicamos duas possíveis medidas de comparação para avaliar os métodos de normalização propostos.

Proporção de spots concordantes

Dado um spot com os valores de intensidade (R_i, G_i) ou ainda (M_i, A_i) , e sua recíproca com intensidades (R'_i, G'_i) ou (M'_i, A'_i) , definimos concordância da seguinte maneira:

$$(M_i, A_i) \text{ e } (M'_i, A'_i) \text{ são concordantes} \Leftrightarrow M_i \cdot M'_i < 0.$$

Medida de concordância por intersecção

Um outro critério para comparação de normalizações, também baseado na idéia de concordância descrita acima, pode ser obtido através de uma métrica de união/intersecção de conjuntos. Em [6], com a intenção de obter uma medida de distância entre comunidades de genes, é descrita a seguinte métrica baseada na união/intersecção:

$$d(A, B) = \frac{\#\{A \cap B\}}{\#\{A \cup B\}} = \frac{\sum_{i,j} \delta_{\alpha_i, \beta_j}}{n + m - \sum_{i,j} \delta_{\alpha_i, \beta_j}}$$

onde A é um conjunto de genes $(\alpha_1, \alpha_2, \dots, \alpha_n)$, B é um outro conjunto de genes $(\beta_1, \beta_2, \dots, \beta_m)$ e

$$\delta_{\alpha_i, \beta_j} = \begin{cases} 1 & \text{se } \alpha_i = \beta_j \\ 0 & \text{caso contrário.} \end{cases}$$

Adaptamos então essa medida de distância para o problema da comparação entre normalizações. Seja C um conjunto de spots fixado. Definimos M como o subconjunto $\{c_i\}$ de C tal que $m_{c_i} < k$, onde m_{c_i} é o valor do log da razão para o spot c_i em uma hibridização fixada. De forma análoga, definimos M' como o subconjunto $\{c'_i\}$ de C tal que $m'_{c'_i} > -k$, onde a hibridização usada como referência deve ser a *recíproca* da hibridização utilizada para achar os spots de M .

Seja então $d_A(M, M')$ a distância entre M e M' , de acordo com a métrica da união/intersecção, quando utilizamos o método de normalização A e $d_B(M, M')$ a mesma medida para quando utilizamos uma normalização B . Se

$$d_A(M, M') > d_B(M, M'),$$

dizemos que o método de normalização A está, sob o critério da métrica união/intersecção, controlando melhor a variabilidade devida à técnica do que a normalização B .

Resultados

lâminas	lowess	splines
rat86/88	15	39
rat85/89	20	38

Tabela 2: Proporção de spots concordantes

lâminas	lowess	splines
rat86/88	0.01	0.21
rat85/89	0	0.20

Tabela 3: Medida de concordância por intersecção

Referências

- [1] CHURCHILL, G. A. Fundamentals of experimental design for cDNA microarrays. *Nat. Gen. Sup.*, v. 32, p. 490-495, 2002.
- [2] CHEN, Y.; DOUGHERTY, E. R.; BITTNER, M. L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, v. 2, n. 4, p. 364-374, 1997.
- [3] YANG, Y. H. et al. *Normalization for cDNA microarray data*. [S.l.], January 2001. Disponível em: <www.stat.berkeley.edu/users/terry/zarray/Html/normspie.html>.
- [4] YANG, Y. H. et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, v. 30, n. 4, 2002.
- [5] FINKELSTEIN, D. B.; GOLLUB, J.; CHERRY, J. M. *Normalization and systematic measurement error in cDNA microarray data*. [S.l.], 2002.
- [6] WILKINSON, D. M.; HUBERMAN, B. A. *A Method for Finding Communities of Related Genes*. Disponível em: <citeseer.ist.psu.edu/546592.html>.

Projeto financiado por:

